

Calibrative approaches to protein solubility modeling of a mutant series using physicochemical descriptors

William F. Long · P. Labute

Received: 8 December 2009 / Accepted: 31 August 2010 / Published online: 15 September 2010
© The Author(s) 2010. This article is published with open access at Springerlink.com

Abstract A set of physicochemical properties describing a protein of known structure is employed for a calibrative approach to protein solubility. Common hydrodynamic and electrophoretic properties routinely measured in the bio-analytical laboratory such as zeta potential, dipole moment, the second osmotic virial coefficient are first estimated in silico as a function a pH and solution ionic strength starting with the protein crystal structure. The utility of these descriptors in understanding the solubility of a series of ribonuclease Sa mutants is investigated. A simple two parameter model was trained using solubility data of the wild type protein measured at a restricted number of solution pHs. Solubility estimates of the mutants demonstrate that zeta potential and dipole moment may be used to rationalize solubility trends over a wide pH range. Additionally a calibrative model based on the protein's second osmotic virial coefficient, B_{22} was developed. A modified DVLO type potential along with a simplified representation of the protein allowed for efficient computation of the second virial coefficient. The standard error of prediction for both models was on the order of 0.3 log S units. These results are very encouraging and demonstrate that these models may be trained with a small number of samples and employed extrapolatively for estimating mutant solubilities.

Keywords Protein solubility · Second osmotic virial · Calibration · Zeta potential

Introduction

In many protein engineering applications one seeks to optimize or control the physical chemical properties of a protein under a given set of conditions. For protein therapeutics, often site directed mutagenesis studies are done to find mutants that have similar activity but have enhanced solubility or solution stability. To a crystallographer, creating mutant series can help optimize protein crystallization conditions thereby eliminating conditions or mutations that do not lead to optimal crystal growth. Often this work is carried out empirically and can lead to unexpected results. A more rational approach, however simple, would immensely help guide the researcher toward their goal.

Bioanalytical laboratories routinely quantitate basic parameters such as size, zeta potential, isoelectric point, the second osmotic virial (B_{22}) and dipole moment. These key parameters aid the understanding of the behavior of colloidal suspensions and proteins. For certain formulations such as injectable drugs one may want to maximize the repulsive forces between particles. This ensures that the suspension does not settle or lead to caking. Alternatively one may want to separate a protein from the mother liquor by minimizing the repulsive forces between particles allowing for efficient flocculation.

The electrostatic properties of a protein play a large role in solution solubility. These properties are governed both by the protein charge distribution and by the nature of the solvent in the formation of the electric double layer. The double layer arises from the presence of surface charges on the protein and its influence on ions at the protein-solvent interface. This results in an increased concentration of ions of opposite charge (counter-ions) near charged surface regions. As one moves further away from the surface, the ion distribution eventually becomes homogeneous as in the

W. F. Long (✉) · P. Labute
Chemical Computing Group Inc., 1010 Sherbrooke St. W, Suite
910, Montreal, QC H3A 2R7, Canada
e-mail: longwf@hotmail.com

bulk solvent. The layer of counter-ions also acts to screen or attenuate surface charges felt further out in solution. Because of these effects, the inner layer of solvent molecules and counter ions surrounding the protein are tightly bound and move in concert. Furthermore there exists an equilibrium distance at which there is a balance between ions that are held by electrostatic forces and those that sheared away due to Brownian motion. This characteristic distance is known as the Debye screening length [1]. This length, κ^{-1} , is typically on the order of tens of Ångströms and depends on the concentration and properties of dissolved salts in solution. At the Debye length these forces may be quantified as measurable electric potential. Theoretically the zeta potential is the average potential, $\varphi(\mathbf{r})$ at the solvation boundary. In the laboratory the zeta potential is determined either by optical means [2] or through electrophoretic band velocity measurements. In its simplest form, the zeta potential, ζ , is related to the electrophoretic velocity, U through the following equation:

$$U = \frac{e\zeta}{\eta} \mathbf{E} \quad (1)$$

where η is the solution viscosity, ε is the dielectric constant and \mathbf{E} is the strength of the applied field. A variety of techniques also exist for estimating ζ from a molecular simulation. Typically these involve the use of the Poisson–Boltzmann (PB) equation [1]. For proteins, boundary integral approaches [3] to solving the PB equation or non-equilibrium molecular dynamics [4] may be used. As practical a rule of thumb, if the measured zeta potential lies within +30 mV to −30 mV this would be indicative of colloidal instability. In this range surface charges on the protein are simply not strong enough for effective protein–protein repulsion.

The second osmotic virial coefficient is also routinely measured electro-optically in the laboratory. This thermodynamic property arises as a higher order term in the equation governing osmotic pressure,

$$\Pi = RTc_P \left(\frac{1}{M_W} + B_{22}c_P + \dots \right) \quad (2)$$

where Π is the osmotic pressure, c_P is the protein concentration, R is the gas constant, T is the absolute temperature and M_W is the protein molecular weight. The second virial coefficient B_{22} indicates the magnitude and direction of the deviations from solution ideality and may be thought of as the effective volume of the protein. This should not be interpreted as a hard sphere volume but one that is influenced by protein–protein electrostatic interactions. By measuring B_{22} at various solution conditions one can gain insight about the underlying effective pair interactions between proteins. A positive value may be interpreted to reflect repulsive interactions whereas a negative value would

indicate the presence of attractive forces. Statistical mechanics also provides a mathematical relationship between the B_{22} and the potential of mean force, U , in an average solvent environment [5]

$$B_{22} = \frac{1}{2M_W\rho V} \int_{\Omega_1} \int_{\Omega_2} \int_0^\infty (1 - e^{-U/kT}) r_{12}^2 dr_{12} d\Omega_1 d\Omega_2 \quad (3)$$

where ρ is the density of the protein, V is its volume, k is the Boltzmann constant and r_{12} is the intermolecular center-to-center distance. For an isotropic sphere, the integral may be interpreted as a Boltzmann weighted volume. However where there is non-uniform charge distribution, it may be more appropriate to think of this as a Boltzmann average over all protein–protein configurations (Ω_1, Ω_2). The portion of the integral corresponding to an overlap where $U \rightarrow \infty$ leads to the excluded volume contribution of the virial coefficient [6].

When estimating B_{22} , often an idealized protein geometry is assumed. Protein molecules are typically treated as uniform spheres with a fixed charge. Although these simplistic models can provide some insight, they may not be reliable near the protein's isoelectric point i.e. when z , the net charge, is zero. If classical DLVO theory [7, 8] were to serve as the electrostatic potential energy function, then the potential U , and thus B_{22} , would be zero. However if charge anisotropy were taken into account, then configurations that lead to energetically favorable interactions between proteins would bias the Boltzmann weighting term in Eq. 3 toward a negative B_{22} . This would indicate protein that would have a higher tendency to agglomerate than what would otherwise be predicted by an isotropic model.

A number of years ago it was shown by George and Wilson that there is a strong correlation between the measured B_{22} and the range of solution conditions that favor protein crystallization [9, 10]. The crystallization success rate was highest if B_{22} falls within a well defined range of -1×10^{-4} to -8×10^{-4} mol mL g⁻². If B_{22} is too large, crystallization rates are slowed due to the dominating repulsive interactions. Conversely if B_{22} is too small, amorphous aggregation can result. It was pointed out that this “crystallization window” corresponds closely to the necessary conditions for the presence of a liquid–liquid immiscibility region in a phase diagram [11, 12]. Further, this region may indicate nucleation via a liquid–liquid phase separation as a first step is a favorable mechanism for crystal growth [12, 13].

Other research efforts have focused on finding a relation between B_{22} and protein solubility in aqueous solutions [13–16]. Finding such a trend is not unexpected as both solubility and the second virial coefficient are governed to a large degree by interactions between protein molecules.

However the relationship between the two is not trivial as solubility is a function of the protein–protein binding energy at specific orientation in the crystal whereas B_{22} is a statistical quantity relating protein–protein interactions in solution over all distance and orientations [13]. The theoretical relation developed by Haas and Wilson has been shown to be in excellent agreement with experimental data for lysozyme and holds over a range solvent pHs, temperature and salt concentrations where the second virial coefficient would be expected to apply [13]. However other studies have shown non-monotonic variation of B_{22} with salt concentration [14]. At high salt concentrations, so called “salting-in” and “salting-out” effects are poorly modeled within the standard DLVO or non-linear Poisson Boltzmann framework.

In this study, our goal is to explore the use of calibrative models of protein solubility at a fixed salt concentration. Such models may prove to be of high value to biochemists who seek insight into which point mutations and pH conditions optimize solubility. In this work we use a crystal structure to provide the necessary information about a protein’s three dimensional charge distribution, although a homology model could be used. Additionally we will explore the possibility of employing a modified DVLO theory that can account for non-isotropic charge distribution in a protein. These ideas will be tested on a series of mutants of Ribonuclease Sa, a well characterized protein for which published solubility data exists [17, 18].

Methods

A protein properties calculator has recently been developed in the Molecular Operating Environment (MOE) [19]. This tool allows one to process one or more proteins, create mutations, and compute a series of protein-specific descriptors. The user can specify the working temperature, pH, solvent dielectric and viscosity and the concentration and nature of the dissolved binary salt. A series of 19 descriptors are output. These range from basic size, eccentricity, volume, and surface area descriptors to electrostatic and transport properties. Most of these properties may be quantified in an analytical laboratory using light scattering, ultracentrifugation and electrophoretic techniques. Although all descriptors were computed only the critical ones actually used in the models are discussed in detail below.

Protein preparation and charge model

Prior to computing the protein descriptors, partial charges were first assigned to all atoms using the AMBER 99 forcefield. If one or more point mutations needed to be made, the appropriate residues sidechains were exchanged

and minimized keeping atom coordinates of all other residues fixed.

For the purposes of determining the protein’s isoelectric point (pI), routines were used that identify titratable groups within the structure and estimate their pK_a values based on the PROPKA method [20]. Given pK_a estimates for each titratable residue, the Sillero and Ribeiro method was then used to estimate the pI of the protein [21]. As this pI estimate is based on pK_a values determined from local 3D residue geometries, it will be referred to as pI_{3D} . For comparison, estimates of pI were also made using fixed group pK_a values. This pI estimate will be referred to as pI_{seq} . The group pK_a values used originate from a recent review [22] of 78 folded proteins (His: 6.6, Lys: 10.5, Arg: 12.3, Asp: 3.5, Glu: 4.2, Cys: 6.8, Tyr: 10.3).

Charges on individual residues were set with the use of the Henderson-Hasselbalch equation [23] given their estimated pK_a values. This ensured a zero net charge at the isoelectric point and allowed for smooth transitions as the pH is varied. It is recognized that charges set in this manner are only appropriate for relative comparisons of similar proteins and are not suitable for molecular mechanical minimizations.

It should also be noted out that in a study of conformational stability of these same proteins [17], the pH of maximum stability for each protein lies near pH 5. The conformational stability measured as a free energy dropped by approximately 50% from its peak value by either raising or lowering the solution pH by 3 units. In light of this, care must be taken when interpreting the physical significance of the descriptors under conditions that may denature or affect the ionization state of the protein.

Zeta potential estimates

From the Gouy-Chapman-Stern model the ζ potential of a charged particle may be determined by estimating the electrostatic potential at the slipping plane in the diffuse layer. It is recognized that although the position of this plane somewhat dependent on the mathematical model and on particle size [24], in this study we are primarily considered with relative changes in ζ potential. We therefore simply assume that this surface is situated at one Debye length, κ^{-1} away from the protein. This characteristic distance is typically on the order of tens of Ångstroms and depends on the ionic strength of dissolved salts in solution, the permittivity of free space, ϵ_0 , and the dielectric constant ϵ_r as well as the absolute temperature, T :

$$\kappa^{-1} = \sqrt{\frac{\epsilon_0 \epsilon_r kT}{2N_A e^2 I}} \quad (4)$$

The other constants in Eq. 4 are Avogadro’s number, N_A , the Boltzmann constant, k , and the charge on

the electron, e . Additionally the ionic strength, I , is defined as

$$I = \frac{1}{2} \sum_i z_i^2 m_i \quad (5)$$

where m is the mass concentration of the i th ionic species and z its valency. For a protein with an inhomogeneous distribution of charges, the electrostatic potential field, $\varphi(\mathbf{r})$ was estimated from a numerical solution of the linearized Poisson–Boltzmann equation. In this study the interior and exterior dielectric constants were fixed at 4 and 78 respectively. The Poisson–Boltzmann equation was solved numerically with the multi-grid preconditioned conjugate residual algorithm in MOE on a $(129)^3$ grid. The grid was constructed so that it encompassed the entire protein and extended one Debye length plus and additional buffer in each Cartesian direction away from the protein's extremities. This ensured a convergent solution over a wide range of solution conditions. The ζ potential was then computed by integrating the potential, $\varphi(\mathbf{r})$ over the solvation boundary:

$$\langle \zeta \rangle = \frac{1}{A_\kappa} \int_{V_\kappa} \varphi(\mathbf{r}) dA \quad (6)$$

where V_κ defines a set of points encompassing the protein and solvation layer and A_κ is its surface area.

Dipole moment

The dipole moment of the protein was computed from Eq. 6 at a given pH-appropriate protonation state:

$$\boldsymbol{\mu} = 4.802 \sum_i^N q_i \mathbf{r}_i \quad (7)$$

where q_i is the partial charge of i th atom and \mathbf{r}_i is the distance in Å from the center of mass. The multiplicative factor was used to convert the dipole moment to units of Debye. AMBER 99 charges calculated in MOE were used as the charge model.

Second virial coefficient, B_{22}

The interaction energies between two proteins were sampled using a 5-axis z - x - z Eulerian gimbal system. This allows one to test any relative configuration. Each protein may be separated by a distance, r_{12} , rotated by an arbitrary longitudinal and azimuthal angle (α , β for protein 1 and θ , φ for protein 2), and rotated by a relative twist angle, γ). As for the potential model used to compute the interaction energies, a variety of forms are typically used and is still the subject of much discussion [6, 13, 14, 16, 25]. However in a study comparing square-well to Yukawa potentials it

was concluded that the relation between the solubility and the second virial coefficient depended very little on the shape of the interaction potential [13]. In light of this, we developed a potential model that was easy to compute using the existing machinery in MOE. Toward this end, the electrostatic component of the interaction energy was computed using the following formula,

$$U_{el} = \frac{e^2}{4\pi\epsilon_0\epsilon_r} \sum_i \sum_j \frac{q_i q_j}{r_{ij}} \quad (8)$$

where i and j are atom indices for proteins 1 and 2 respectively, q is the partial charge on a given atom and r_{ij} is the relative distance between atoms i and j . Unfortunately Eq. 8 does not account for inter-particle charge screening effects due the presence of dissolved salts in the medium. In order to account for this, numerical solutions of the Poisson Boltzmann equation would be normally be required for all configurations. However such an approach would simply be too time prohibitive. There exists however a linearized solution to the Poisson Boltzmann equation for two spheres of radius α with uniform surface charge z_0 placed in a dielectric medium separated by a distance r_{12} . This approximation is the framework of Derjaguin, Landau, Verwey and Overbeek (DLVO) theory [7, 8]. The electrostatic component of this classical theory is,

$$U_{DLVO}(r_{12}) = \frac{(z_0 e)^2 \exp(2\kappa\alpha) \exp(-\kappa r_{12})}{4\pi\epsilon_0\epsilon_r (1 + \kappa\alpha)^2 r_{12}}. \quad (9)$$

If a system of two particles both having a charge z_0 are considered, then by inspection it may be seen that Eq. 9 closely resembles Eq. 8 multiplied by a scaling factor that accounts for dielectric screening. For large separation distances, r_{12} , Eq. 8 may be modified to:

$$U(r_{12}) \approx \frac{e^2}{4\pi\epsilon_0\epsilon_r} \frac{\exp(2\kappa\alpha) \exp(-\kappa r_{12})}{(1 + \kappa\alpha)^2} \sum_i \sum_j \frac{q_i q_j}{r_{ij}} \quad (10)$$

This approximation will serve as the electrostatic model. For small distances ($r_{12} \leq \alpha$) inside the exclusion volume, a hard sphere potential will be used.

Additionally there exists several forms of the multiplicative constant outside of the configuration integral in Eq. 3 [6, 13, 14]. In general this factor depends on protein molecular weight, density and volume and a normalization constant. Since we will employ a calibrative approach to estimating the solubility from B_{22} , we are less concerned this factor and allow it to be a free parameter, a , in our model. We leave the divisor $2M_w\rho V$ in place to account for the units and to ensure proper scaling. The Boltzmann volumetric weighting factor in Eq. 3 also presents a challenge when one encounters either a very favorable (quasi-docked) protein–protein configuration or a low energy sample not in the training set. Because of this we employ a

modified potential $b[U(r) - c\langle U_0 \rangle]$ where $\langle U_0 \rangle$ is an estimate of the average minimum interaction energy over all samples under consideration and b and c are free scaling factors. The minimum interaction energy $\langle U_0 \rangle$ is needed especially when dealing with protein solutions that are not infinitely dilute or when solving Eq. 3 numerically. At a typical working protein concentrations for crystallization ($\sim 15 \text{ mg mL}^{-1}$), the average maximum protein–protein separation distance is on the order of 115 \AA and gives a rough indication of the required upper limit of integration along r_{12} . Although interaction energies beyond this point may be non-zero, they will not contribute significantly to the Boltzmann average. We therefore choose $c\langle U_0 \rangle$ as our reference point. For our modified potential, the scaling factors b and c can be seen to account for discrepancies between the theoretical and actual exclusion volume, solution conditions and choice of energy scale. The functional form of B_{22} becomes:

$$B_{22} = \frac{a}{2M_w \rho V} \int_{\Omega_1} \int_{\Omega_2} \int_0^\infty (1 - e^{-b(U - cU_0)/kT}) r_{12}^2 dr_{12} d\Omega_1 d\Omega_2. \quad (11)$$

Continuing toward our ultimate goal of relating the second virial to solubility a survey of the literature has revealed a few thermodynamical and empirical forms [13, 15, 16]. An equation developed by Haas et al. provides us with a mathematically stable functional form for the fitting procedure [12]. Writing the solubility in terms of B_{22} we use the following equation:

$$S = 55.55 M_w \left(\frac{1 - \rho M_w B_{22}/4}{d} + 1 \right)^{-N/2} \quad (12)$$

where d is a parameter which depends on the anisotropy, N is the coordination number and S is the solubility in mg mL^{-1} . Typically N is between 3 and 6 and d is on the order of 0.001. The coordination number is reported to be related to be number of nearest neighbor protein molecules in the crystal.

Finally given a set of experimental solubilities, conditions and a crystal structure for the training compounds, and an estimate of $\langle U_0 \rangle$, we fit the optimal parameters in Eqs. 11 and 12 that minimize $(\log S - \log \hat{S})^2$ where \hat{S} is the estimated solubility. Non-linear optimization functions in MOE based on the steepest decent method were used to perform the fitting.

Reduced protein representation

With a fully atomistic representation of the protein, the time required to evaluate a large number interaction energies makes the above computation of B_{22} impractical.

Although there are many ways to reduce computation times, a convenient solution was found by replacing the atoms of each amino acid residue with two that preserve the net residue charge and the residue dipole moment. In this reduced representation, two “atoms” were placed at the centers of negative and positive charge respectively. The radius α , of this reduced representation was taken to be that of a sphere that minimally encompassed the reduced set of atoms plus a small buffer. With this approach, the protein’s net charge and higher moments are retained. Additionally the potential energy machinery in MOE could be employed without modification.

Modeling

The crystal structure of Ribonuclease Sa (PDB code: 1RGG) was submitted to the protein properties calculator. Descriptors were computed at the pH values shown in Fig. 1 for the wild type, 3K (D17K + E41K + D1K), and 5K (D17K + E41K + D1K + D25K + E74K) mutants. The dielectric medium was specified to be a 10 mM NaCl solution at 298 K. The solvent viscosity was set to that of water, 0.89 cP . For the intermediate calculations required to estimate B_{22} , a total of 351,000 protein–protein configurations were sampled. This is comprised of 26 separation distances from 0 to 104 \AA , and 30 evenly sampled sets of azimuthal and longitudinal angles for each protein and 15 twist angles. A total computation time of 14 h was required for the 24 samples on a 1.8 GHz Intel dual core system.

After the descriptor calculation was complete, a fitting procedure using experimental solubilities from the training

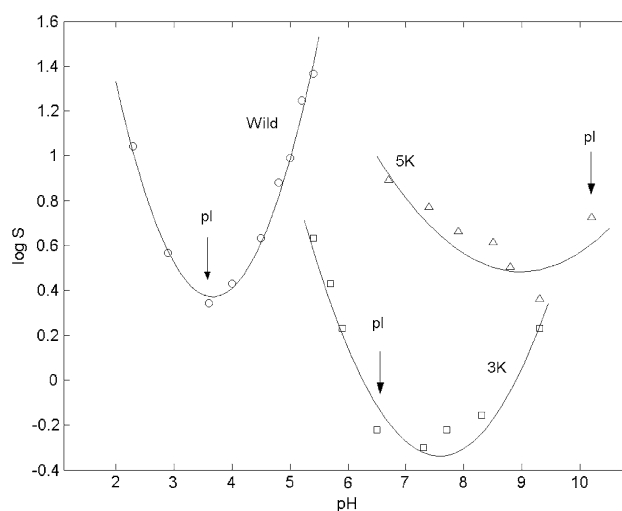


Fig. 1 Measured solubilities as a function of pH for RNase Sa (circles), 3K (squares) and 5K variants (triangles). The lines are meant as a guide only. This figure is reconstructed from Ref. [13]

set allowed for the completion of the descriptor table with estimates of B_{22} and $\log \hat{S}$ for all samples. For these calculations, we assumed a density of 1.36 g cm^{-3} for the protein.

Additional pI values were estimated for the 2K (D17K + E41K) and 4K (D17K + E41K + D1K + D25K) variants for comparison with experimental values.

Results and discussion

A set of protein descriptors was computed using the protein properties calculator for Ribonuclease Sa and four lysine mutants at solution pHs where published pI and solubility data exists [17]. The wild type protein contains 96 residues, has a pI of 3.5 and contains no Lys residues. A set of variants was created by replacing highly exposed surface Asp and Glu residues with Lys. This is of interest given that the natural substrate for the ribonucleases is negatively charged, the net charge on the enzyme might influence its steady-state kinetics [17]. As shown in Table 1, the isoelectric points for the 2K, 3K, 4K and 5K variants show an increasing shift toward higher pH. The estimated isoelectric points, pI_{seq} and pI_{3D} , are based respectively on fixed group pK_a s and those determined from local 3D residue geometries. With this small sample, the average error between the estimated and experimental values is 0.66 pH units for pI_{seq} and 0.45 pH units for pI_{3D} . In this set, the pI estimate of the 5K variant appears to have the largest error of approximately 1 pH unit. This may be due to the fact that RNase Sa has eight Tyr that are largely buried having experimental pK_a values greater than 11. This is much higher than the pK_a of 10 that is assumed in the calculation of pI_{seq} . However with most theoretical determinations of pI, is it a given that the results are accurate to within ± 1 pH unit of the experimental value [17, 26]. Although these estimates based on pI_{3D} look very encouraging for this set, we have not yet performed any larger scale studies validating the method and have used these estimates simply to set the point of zero change for charge based descriptors.

The protein properties calculator also returns pK_a estimates of the individual ionizable amino acids given their

relative 3D environment in a protein using the PROKPA method. Estimates for RNase Sa and its 5K variant are listed in Table 2 along with of their respective measured values [27]. For this system the pK_a s may be estimated to within ± 1 pH unit of the measured value. It is interesting to note that the pK_a estimates for each ionizable group are almost identical between the wild type and the modeled 5K variant. Only the pK_a of Glu78 was found to decrease by 0.1 pH unit. However a different picture emerged when the group pK_a estimates were made on the known crystal structure of the 5K variant (PDB code 3A5E). In this case the estimated pK_a s of eight residues were different.

In order to understand the difference between these modeled 5K variant and its known crystal structure, two types of alignments were performed. The first was an alignment on the backbone alpha carbons. Since there was only a the 0.37 Å RMSD between the two structures, it may be concluded that the surface mutations do not influence the overall protein structure. When the second alignment was done using the heavy atoms of the five lysine sidechains, the RMSD was found to be 1.9 Å. Given that the level of variation, the scoring function employed in PROPKA used to shift the pK_a from its baseline value seems quite sensitive to sidechain configuration. This scoring function attempts to account for desolvation, hydrogen bonding and buried charge–charge interactions [20]. Since there is such a difference in the predicted shift in pK_a between the modeled 5K mutant and the known crystal structure, this would indicate that the surface lysines in the model are not forming as many hydrogen bonds with nearby residues. This is in fact observed by inspection. In light of this result, it may be better to first perform a short molecular dynamics run on the sidechain atoms in order to generate an ensemble of conformers then compute a Boltzmann-weighted average of the predicted group pK_a s. The only drawback to this approach is the lengthy computation time.

Solubility modeling with simple electrostatic descriptors

For an initial model, data from the wild type and 3K variant served as the training set. The remainder from the 5K variant served as the test set used for model assessment. This calibration set was chosen to cover the entire dynamic range of solubilities. An initial study was done to determine which of the 19 descriptors were correlated with the experimentally available solubility values in the training set. For the charge based descriptors z (net charge) and ζ (zeta potential), their square values were used. This was done because, to a first approximation, the solubility of a protein is known to be proportional to the square of the net change on the protein [28]. The characteristic parabolic shape shown in the solubility profiles in Fig. 1 clearly

Table 1 Isoelectric points of RNase Sa and the 2K, 3K, 4K and 5K variants

RNase Sa variant	Experimental pI	pI_{seq}	pI_{3D}
Wild type	3.5	3.9	3.5
2K	4.6	4.8	4.5
3K	6.4	6.2	6.6
4K	8.4	7.1	8.0
5K	10.2	8.4	9.3

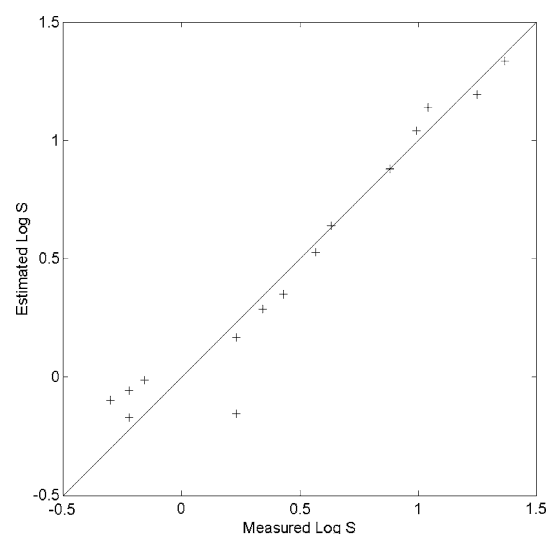
Table 2 pK_a values for RNase Sa and its 5K mutant

Residue	RNase Sa measured	RNase Sa estimated	5K variant measured	5K model estimated	5K crystal estimated
Asp01	3.44	2.6	Lys	Lys	Lys
Asp17	3.72	3.9	Lys	Lys	Lys
Asp25	4.87	3.8	Lys	Lys	Lys
Asp33	2.39	1.5	2.12	1.5	2.1
Asp79	7.37	5.2	6.06	5.2	5.1
Asp84	3.01	1.1	2.98	1.1	1.1
Asp93	3.09	2.4	3.00	2.4	2.4
Glu14	5.02	4.1	2.83	4.1	4.1
Glu41	4.14	4.5	Lys	Lys	Lys
Glu54	3.42	2.5	2.26	2.5	4.2
Glu74	3.47	4.5	Lys	Lys	Lys
Glu78	3.13	2.9	3.02	2.8	2.6
His53	8.27	6.3	7.39	6.3	7.2
His85	6.35	7.0	6.09	7.0	6.4
Tyr30	11.3	10.1	n.d.	10.1	10.0
Tyr49	10.6	10.0	n.d.	10.0	9.9
Tyr 51,52,55,80,81,86	>11.5	11.3 (avg)	n d.	11.3 (avg)	11.3 (avg)

demonstrates this trend: proteins are usually least soluble near their isoelectric points and their solubility increases when either the pH is raised or lowered. Furthermore the descriptor values were autoscaled to avoid biasing and the log of the solubility (log *S*) was used throughout.

Working with the 17 data points in the training set, a stepwise multi-linear regression procedure was used for variable selection [29]. With this procedure candidate variables are included in the model one by one. This process continues until successive models are deemed not to be statistically by means of an *F*-test at the 95% confidence level. Using this method it was found that ζ^2 and μ (dipole moment) were most correlated to log *S*. It is interesting to note that although z^2 is also correlated to log *S* ($r^2 = 0.77$), it was not selected with the stepwise procedure given that ζ^2 is slightly more correlated ($r^2 = 0.83$). It was also found that after removing two of the 3K samples (at pH 5.4 and 5.7) the standard error of prediction for the training set decreased significantly from 0.24 to 0.15. We therefore constructed our models without them. This leads to a simple relation which will be referred to a Model 1: $\log S = 2.50 + 0.00274\zeta^2 - 0.00266\mu$ ($r^2 = 0.968$, $R^2 = 0.937$). As shown in Fig. 2, a plot of experimental log *S* versus estimated log *S* for the calibration set demonstrates that the residues are homoscedastically distributed about the line of identity.

A more informative view of the results demonstrating the accuracy of the model is shown in Fig. 3. Here the estimated and experimental solubilities are plotted as a function of pH. The standard error of prediction on the test set (seven 5K samples) was 0.393. Results indicate that the

**Fig. 2** Plot of the measured versus estimated log *S* on the training set. A line of identity is plotted along the diagonal

model reproduces the characteristic parabolic trend in solubility about the pI in the wild type and shows a minimum point with the 3K and 5K variants. The model also reproduces the non-linear progression in solubility as the number of positively charged Lys groups is increased. However the point of minimum solubility for the 3K variant predicted by the model is shifted 0.5 pH units lower than what appears experimentally but remains close to the observed pI. From the reported experimental conditions [17] it should be noted that the buffer composition is varied according to the pH which may lead to some uncertainty as where the true minimum solubility actually lies. The

conclusions that may draw however are clear: the 3K variant is predicted to be less soluble than the wild type and the 5K slightly more soluble than the wild type at their respective isoelectric points. In spite of these results there exists a fundamental problem as the chemist would need training data that is likely to cover the full dynamic range of solubilities ensuring the model be used in an interpolative manner. This calibrative procedure has little utility since the chemist might only have solubility data of the wild type and wants to know in advance if it is worth expressing and testing a given mutant!

A more realistic approach to solubility modeling would be to train the model using data from a single protein and employ it in an extrapolative sense to estimate solubilities of similar proteins at a desired pH. Toward this goal, the above process was repeated using five calibration points from the wild type. A two parameter model based on ζ^2 and μ lead to the following regression equation (Model 2): $\log S = 2.68 + 0.00246\zeta^2 - 0.002805\mu$ ($r^2 = 0.9985$, $R^2 = 0.9971$, SEP = 0.33). The solubilities of the wild type are estimated very well as is shown in Fig. 4, however for the 3K and 5K variants, only the broad trends are captured. It is interesting to note that even with much fewer training points, the calibration coefficients for Model 1 and Model 2 are similar. This type of model is more error prone but more realistic in how it would be used. In this case, the chemist would get a semi-quantitative picture that around their respective isoelectric points, the 3K is less soluble than the wild type and the 5K is slightly more soluble.

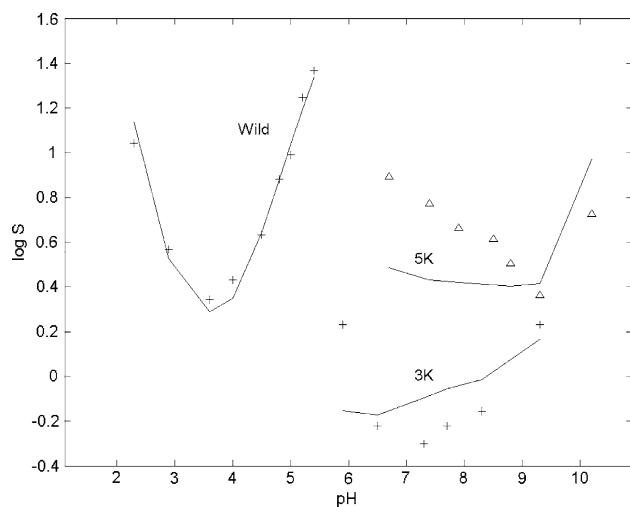


Fig. 3 Plot of the measured and estimated $\log S$ as a function of pH. Estimated values were determined from Model 1. Training data from the wild type and 3K variant is shown as crosses. Experimental solubilities in the test set (5K variant) are shown as triangles. Estimated values are shown with lines

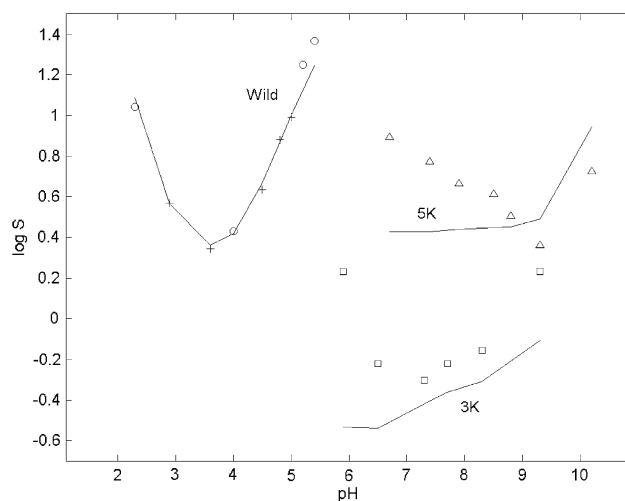


Fig. 4 Plot of the measured and estimated $\log S$ as a function of pH. Estimated values were determined from Model 2. Training set data is shown as crosses. Experimental solubilities belonging to the test set are shown as either circles (wild type), squares (3K variant) or triangles (5K variant). Estimated values are shown with lines

Physical interpretation of solubility models based on simple electrostatic descriptors

Plots of the estimated zeta potential and dipole moments as a function of pH are shown in Figs. 5 and 6 respectively. The zeta potential plot is a type protein titration curve with ζ playing the role of an indicator. This may be used to approximate the isoelectric point (pI)—the point where the electrophoretic mobility is zero. If the protein were an isotropic charged sphere of radius a , then at the pI both the net charge and ζ would be zero. This condition arises naturally from the solution of the Poisson Boltzmann equation for an isotropic sphere as $z = 4\pi\epsilon_0 a(1 + \kappa a)\zeta/e$. If this relation holds more generally, then one can make comparisons between protein solutions over a range of pHs at same ionic strength. With this in mind, it may be seen from Fig. 5 that pI of the mutant series generally increases with the number of Lys groups. Additionally, the slopes of the linear regions near the pI for the wild type and 3K variants protein give an indication of the steepness of their corresponding parabolic shaped solubility curves in Fig. 1. This is because the protein's relative solubility is typically proportional to square of the net charge. However for the 5K variant, at 2 pH units lower than the pI there is a gradual downward slope in ζ whereas the slope is more pronounced at 2 pH units higher than the pI. In this case one would expect a non-symmetric trend in solubility about the pI which is indeed observed.

A plot of the estimated protein dipole moments as a function of pH shown in Fig. 6 demonstrates a different set of trends. In this case the curve shapes of the wild type and variants are similar but the magnitude of the dipole

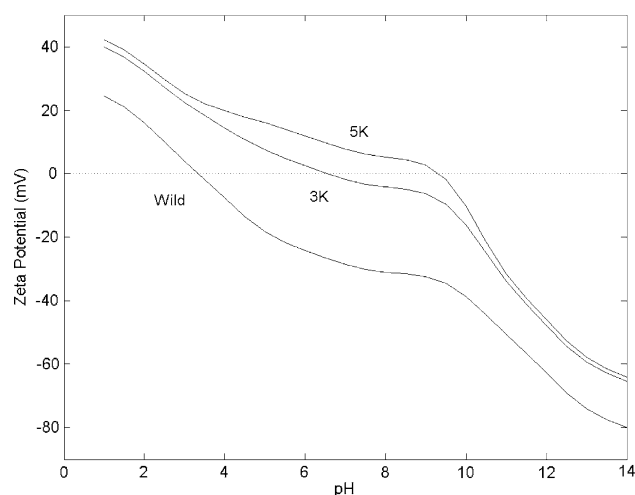


Fig. 5 Plot of the estimated zeta potential as a function of pH for the wild type, 3K and 5K variants

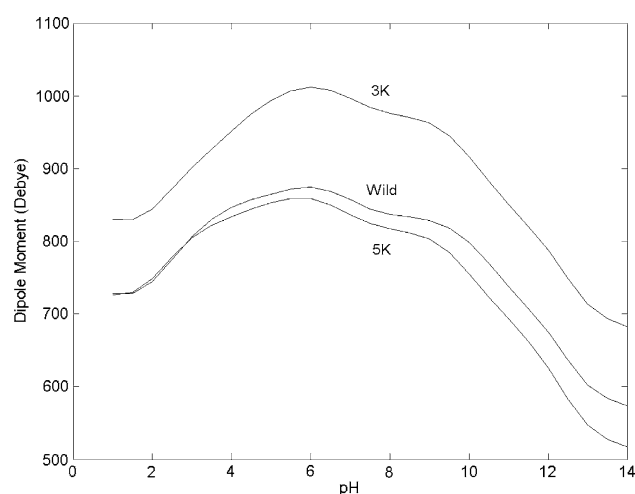


Fig. 6 Plot of the estimated dipole moment as a function of pH for the wild type, 3K and 5K variants

moment mirrors the trend in solubility at their respective pIs. It should be mentioned that since we are concerned with relative changes in dipole moment for a mutant ladder series, the use AMBER 99 charges is sufficient. For absolute measures of dipole moment other charge models may be more appropriate [30]. Generally a larger dipole moment represents greater separation of charge resulting in a more polarized system. This may lead to increased protein–protein interactions by attraction of oppositely charged regions. When this occurs solubility may decrease due to flocculation.

Solubility modeling with the second osmotic virial

A solubility model based on the expression for the second osmotic virial (Eq. 11) together with a functional form

relating it to solubility (Eq. 12) was also developed. A reduced protein configuration was used in the sampling 351,000 interaction energies needed for the configurational integral at the heart of the B_{22} estimate.

Model training was also made using samples from the wild type. One of the issues encountered when allowing each parameter a , b , c , d and N to vary is that more training data was required. Even using all available samples, it was found that the d and N parameters were too interdependent to allow for rapid convergence. In studies with lysozyme, it is reported [13] that a good fit between experimental data and Eq. 12 was obtained for $d = 0.01$ and $N = 4$. In the absence of any a priori B_{22} data to suggest otherwise, we held d and N fixed at these values. Therefore Eq. 12 simply serves as a generic mapping function relating B_{22} to the solubility. The fitting process was done with the same five calibration points from the wild type as used previously.

The best fit parameters a , b and c values were found to be -0.166 , 0.23 and 0.136 respectively and the standard error of prediction was 0.201 . The results for Model 3 shown in Fig. 7 demonstrate that the model captures the broad trends in solubility in the test set. The solubilities of the wild type are again determined quite well, and the overall trends in solubility for the 3K and 5K are as good as those predicted in Model 2. Using the best fit a , b and c values, one may estimate which proteins/conditions fall into crystallization window i.e. the range where B_{22} is between -1×10^{-4} and -8×10^{-4} mol mL g^{-2} . It is recognized however that since we employed a fixed, generic mapping function relating B_{22} to the solubility, the corresponding “solubility window” is rather approximate. The use of experimental B_{22} data however may be used to

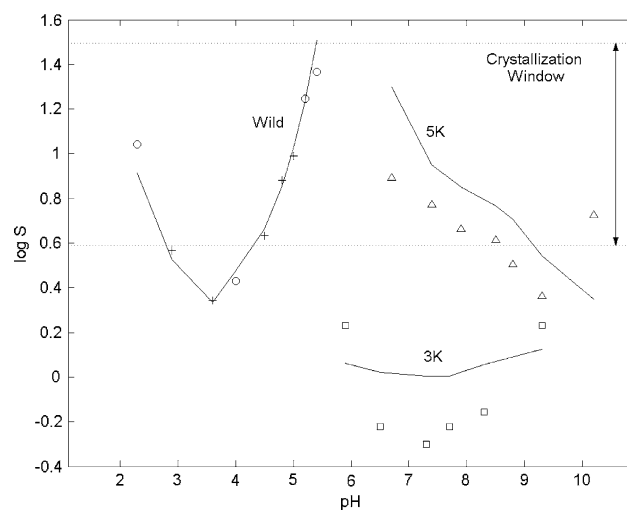


Fig. 7 Plot of the measured and estimated $\log S$ as a function of pH. Estimated values were determined from Model 3. Training set data is shown as crosses. Experimental solubilities belonging to the test set are shown as either circles (wild type), squares (3K variant) or triangles (5K variant). Estimated values are shown with lines

set more appropriate values for d and N . The solubility window shown in Fig. 7 suggests that the wild type and 5K variants might have a greatest chance of being crystallized at ± 1 pH units on either side of their respective isoelectric points. However the 3K variant appears to be too insoluble under the current solution conditions.

Additionally, preliminary tests were performed by allowing d vary between 0.01 and 0.035 and N from 3 to 4. As no significant change in the error of prediction was found in this range, the values of the fitting parameters a , b and c adjusted to compensate for imprecise knowledge of d and N . This indicates that one might not need to explicitly include d and N in the fitting procedure as they are relatively insensitive. From a practical standpoint this may be advantageous as fewer calibration samples would be required leaving only the need to pick from amongst a small set of pre-determined generic mapping functions that best suits a given protein family.

Conclusions

In this study we have developed simple calibrative approaches to estimating protein solubility as a function of pH using a set of descriptors calculated from its 3D structure. The models are purely electrostatics-based and other factors that may affect solubility such as hydrophobic interactions and salting out at high salt concentration are not accounted for. However even with this limitation, solubility models may be still be constructed when estimates for a mutant series are sought. A solubility model of Ribonuclease Sa based on zeta potential and dipole moment trained on wild type data only demonstrate this. Additionally the approach allows one to rationalize solubility trends. Another solubility model based on the second virial coefficient also provides information of interest to crystallographers who wish to determine whether a given mutant may likely have a chance of being crystallized. These results are encouraging given that the models are extrapolative, based on a range of different theories and trained on experimental data that is not highly accurate by its nature. The protein descriptors employed in this work may be computed very efficiently and can provide the analyst with estimates of critical parameters that take much longer to determine experimentally. In the future we wish to extent the methods to non-globular proteins and investigate the effect of amino acid side chain conformation on model stability.

Open Access This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

References

- Henderson D, Wasan DT (1999) In: Hsu JP (ed) Interfacial forces and fields: theory and applications. Marcel Dekker, New York
- Tscharnuter WW (2001) Appl Opt 40:3995
- Chae KS, Lenhoff AM (1995) Biophys J 68:1120
- English N, Long WF (2009) Physica A 388:4091
- Landau LD, Lifshitz EM (1980) Statistical physics, vol 5, 3rd edn. Pergamon, Oxford
- Neal BL, Asthagiri D, Lenhoff AM (1998) Biophys J 75:2469
- Derjaguin B, Landau L (1941) Acta Physico Chimica URSS 14:633
- Verwey EJW, Overbeek JTG (1948) Theory of the stability of lyophobic colloids. Elsevier, Amsterdam
- George A, Wilson WW (1994) Acta Crystallogr D 50:361
- George A, Chiang Y, Guo B, Arabshahi A, Cai Z, Wilson WW (1997) Methods Enzymol 276:100
- Lekkerkerker HNW (1997) Physica A 244:227
- Haas C, Drenth JJ (1998) J Phys Chem B 102:4225
- Haas C, Drenth JJ, Wilson WW (1999) J Phys Chem B 103:2808
- Allahyarov E, Loewen H, Hansen JP, Louis AA (2003) Phys Rev 67:E5140
- Ruppert S, Sandler SI, Lenhoff AM (2001) Biotechnol Prog 17:182
- Guo B, Kao S, McDonald H, Asanov A, Combs LL, Wilson WW (1999) J Cryst Growth 196:424
- Shaw KL, Grimsley GR, Yakovlev GI, Makarov AA, Pace CN (2001) Protein Sci 10:1206
- Trevino SR, Scholtz JM, Pace CN (2007) J Mol Biol 366:449
- Molecular Operating Environment version 2008.09, Chemical Computing Group Inc. Montreal, Canada
- Li H, Robertson AD, Jensen JH (2005) Proteins 61:704
- Sillero A, Ribeiro JM (1989) Anal Biochem 179:319
- Pace CN, Grimsley GR, Scholtz JM (2009) J Biol Chem 284:13285
- Cameselle JC, Ribeiro JM, Sillero A (1986) Biochem Educ 14:131
- Yoon BJ (1991) J Colloid Interface Sci 142:575
- Petsev DN, Vekilov PG (2000) Phys Rev Lett 84:1339
- Patrickios CS, Yamasaki EN (1995) Anal Biochem 231:82
- Laurents DV, Huyghues-Despointes BMP, Bruix M, Thurlkill RL, Schell D, Newsom S, Grimsley GR, Shaw KL, Trevino S, Rico M, Briggs JM, Antosiewicz JM, Scholtz JM, Pace CN (2003) J Mol Biol 325:1077
- Tanford C (1961) Physical chemistry of macromolecules. Wiley, New York
- Draper N, Smith H (1981) Applied regression analysis, 2nd edn. Wiley, New York
- Felder CE, Prilusky J, Silman I, Sussman L (2007) J Nucl Acids Res 35:W512